Creativity of images: using digital consensual assessment to evaluate mood boards

By: Charles Freeman, Sara Marcketti, and Elena Karpova

Freeman, C., Marcketti, S., & Karpova, E. (2017). Creativity of images: Using digital consensual assessment to evaluate mood boards. Fashion and Textiles, 4(17), 1-15. DOI 10.1186/s40691-017-0102-4

Abstract:

Mood boards are used frequently in design and product development as well as in academic courses related to fashion design. However objectively evaluating the creativity of fashion design mood boards is often difficult. Therefore, the purpose of this investigation is to examine reliability of a digital consensual assessment instrument measuring creativity, using expert (from related domains) and non-expert raters (students). Creativity measures were compared with the mood board themes to further investigate any relationships between mood board types and the consensual assessment. Independent samples t test comparing group means indicated expert raters evaluated the mood boards significantly higher in creativity than the non-experts, $t(99) = -6.71$, $p < .001$, (95% CI −.57, −.29), while Pearson correlation results indicate a significant relationship between the two groups of raters, $r(50) = .33$, $p < .01$. ANOVA results for all raters indicated a significant difference between the five subject matter categories; $F(4, 95) = 4.64$, $p < .005$. Overall, expert and non-expert raters reported significant reliability, which further supports prior research using consensual assessment for creativity measures across domains.

Keywords: Creativity | Mood boards | Consensual assessment | Digital creativity evaluation

Article:

***Note: Full text of article below

Freeman *et al. Fash Text (2017) 4:17*
DOI 10.1186/s40691-017-0102-4

Fashion and Textiles

RESEARCH Open Access

Creativity of images: using digital consensual assessment to evaluate mood boards

Charles Freeman[1*], Sara Marcketti[2] and Elena Karpova[2]

*Correspondence: cfreeman@humansci.msstate.edu
[1] Mississippi State University, Mississippi State, MS, USA
Full list of author information is available at the end of the article

Abstract

Mood boards are used frequently in design and product development as well as in academic courses related to fashion design. However objectively evaluating the creativity of fashion design mood boards is often difficult. Therefore, the purpose of this investigation is to examine reliability of a digital consensual assessment instrument measuring creativity, using expert (from related domains) and non-expert raters (students). Creativity measures were compared with the mood board themes to further investigate any relationships between mood board types and the consensual assessment. Independent samples *t* test comparing group means indicated expert raters evaluated the mood boards significantly higher in creativity than the non-experts, $t\,(99) = -6.71$, $p < .001$, (95% CI −.57, −.29), while Pearson correlation results indicate a significant relationship between the two groups of raters, $r\,(50) = .33$, $p < .01$. ANOVA results for all raters indicated a significant difference between the five subject matter categories; $F\,(4, 95) = 4.64$, $p < .005$. Overall, expert and non-expert raters reported significant reliability, which further supports prior research using consensual assessment for creativity measures across domains.

Keywords: Creativity, Mood boards, Consensual assessment, Digital creativity evaluation

Introduction

Fashion mood boards are fundamental tools in design and merchandising fields. Using color, texture, image, form, and sometimes objects, mood boards bring to visual life a feeling or sentiment (Garner and McDonagh-Philp 2001). Mood boards cover a wide range of visual representations from concept boards to sales boards, each having a specific use: from inspiration for designers to product communication for merchandisers. Research examining the evaluation of mood boards, namely creativity and expression of thematic elements, is lacking, perhaps because of the subjective nature often present when representing through visual means a feeling or idea (McDonagh and Storer 2004). Overall, creativity assessment in apparel and textile research is relatively limited, yet few researchers dispute the importance of creativity in the domain.

Conceptual mood boards are typically non-product specific and include a range of images creatively representing a theme or idea. These images set an overall feel for the

project and are utilized as a creative source of inspiration or the exploration of project/product ideas (Cassidy 2008). Because of their use often in the early stages of the design process (Lucero 2012), non-detailed guidelines regarding how they are created are typical, with even less objective directions regarding their evaluation provided. While the depth and quality of the board is largely dependent on the creativity (unspecified nor assessed) in the selection of images, there exists little research regarding how the creativity of the mood boards may be evaluated. As such, this research applied a specific tool, the Consensual Assessment Technique (CAT) to the evaluation by both experts and non-experts of mood boards.

Questions linger about the domain specificity of creativity assessment and methods of identifying and classifying experts. Plucker and Runco (1998) stated the uselessness of single predictive creativity measures, such as creativity tests, and agreed with Hennessey and Amabile (1988) of using an overall assessment of product output, measured using consensual assessment. In prior research on expert and non-expert creativity assessment, the domains were limited to those ranking higher on Simonton's (2009) hierarchy of domains, such as poetry, and creative writing. In recent years, fashion has slowly climbed from a utilitarian artifact with functional appeal to creative applied arts, with exhibitions in the finest museums around the world. With this rise to new creative heights, there exist a deficit in understanding and evaluating creative outputs, starting with the concept mood board stage, in the current research. In addition, much of the prior research using verbal stimuli resulted in effective creativity evaluation, however assessment of figural stimuli is limited when comparing expert and non-expert assessments of creativity (Baer et al. 2004; Kaufman et al. 2009). Despite the deficit in understanding and evaluating creativity, much of the prior research conducted includes both expert and non-expert groups to provide assessment beyond the scope of technical quality (Kaufman and Baer 2012). Therefore, the purpose of this investigation is to examine to reliability and validity of the CAT when using expert (faculty) and non-expert raters (students) to evaluate the creativity of images centered on a limited selection theme of a fashion mood board. Additionally, results from the CAT will be compared with mood board themes to further investigate any relationships between themes/ideas and consensual assessment.

Literature review

Defining creativity

In beginning research into the relationship between creativity and fashion design mood boards, the initial complexity of this relationship is the various definitions of creativity offered across domains. As an example, the following definitions taken from four major researchers in creativity; J.P. Guilford, Theresa Amabile and Howard Gardner respectively, show the various components and layers of creativity definition:

> *"a creative pattern is a manifest in creative behavior, which includes such activities as inventing, designing, contriving, composing, and planning. People who exhibit these types of behavior to a marked degree are recognized as being creative" (Guilford 1950, p .444)*

> *"a product or response will be judged creative to the extent that (a) it is both novel and appropriate, useful, correct, or valuable response to the task at hand and (b) the task is heuristic rather than algorithmic" (Amabile 1983a, p. 33)*

> *"the creative individual is a person who regularly solves problems, fashion products, or defines new questions in a domain that is initially considered novel but that ultimately becomes accepted in a particular cultural setting" (Gardner 1993, p. 35).*

As can be inferred from these various definitions, there are multiple levels of creativity definition, with a few recurring thematic concepts. Generally, a person or product is considered to be creative based on novelty to the domain or field of work. In addition to novelty, the creative process and output evaluation and acceptance are commonplace in creativity definitions. Since the early 1950s research in the area of creativity has led to acceptance of the following four areas of focus (Four P's) in creativity studies: (1) person, (2) process, (3) product and (4) press (environment) (Kaufman and Baer 2012). Initially, much of the research completed focused on the person and personality, but contemporary theoretical models exhibited the influence of the other three major concepts and show the relevance of evaluation and assessment of the creative product in relation to fashion designers and the fashion design process.

Fashion mood boards

A powerful communication tool, the fashion mood board comprises incongruent and seemingly un-related images which when tactfully selected provide complementary support in creating a coherent visual message discussing a specified theme or idea (Boyes 1998). Yet, the currently available research specific to mood boards is lacking and limited to a few studies within the past few years. Mood boards are a tool used to visually communicate information in fashion-related and consumer product industries (Cassidy 2011). Primary uses include bringing together images and ideas into an aesthetically and creatively impactful workspace, for a specific purpose. From an academic standpoint the inclusion of mood boards on projects and within courses provides training for industry practice, while providing students with the opportunity to communicate their ideas visually. Within this space, designers select and arrange images, artifacts, colors, fabrics, etc. in a calculated and planned effort to link content to a particular theme or idea. However, there exists limited training or skill development related to mood boards in academia, especially in relation to mood board evaluation and improvement, as well as the role in their role in the design process (Cassidy 2011).

Categorically, Cassidy (2008) defined four types of mood boards and their specific uses as indicated following. The mood boards used are examples created within the past 5 years by students for a course project. While these boards may not score high on creativity assessment, boards were purposely selected as evidence specific to the various categories identified by Cassidy (2011).

Category 1 boards (Fig. 1) are focused on target market identification, including lifestyles, socio-cultural factors, demographics, personal and cultural values/norms, and links between company and target market values. These boards are often labeled as lifestyle, target, customer, or consumer profile boards.

Fig. 1 Example of Category 1 concept board

Category 2 (Fig. 2) is more conceptual in comparison and utilized during the initial stages of the product design process. Often focused on a feeling or exploration of ideas, board content is typically non-specific and derived from target market lifestyles. Through visual communication, these boards provide clarity and vision for a thematic concept derived from design briefs. Most often these are referred to as concept, idea, inspiration, story or style boards. These boards are conceptual in content and interpretation and therefore will be the focus of this investigation.

Fig. 2 Example of Category 2 concept board

Category 3 boards (Fig. 3) enhance and sharpen the concepts and ideas developed in category 2 boards. Specificity is more apparent at this level and includes a variety of boards with clearly defined purpose: color, fabric, style, trend, samples and forecasting. Much of the work utilizing boards at this stage are focused on refining the product identity and message, through visual representation and comparison.

Category 4 boards (Fig. 4) are the most refined and professional boards, as they represent the final product line carried to clients and consumers. Often used by marketing

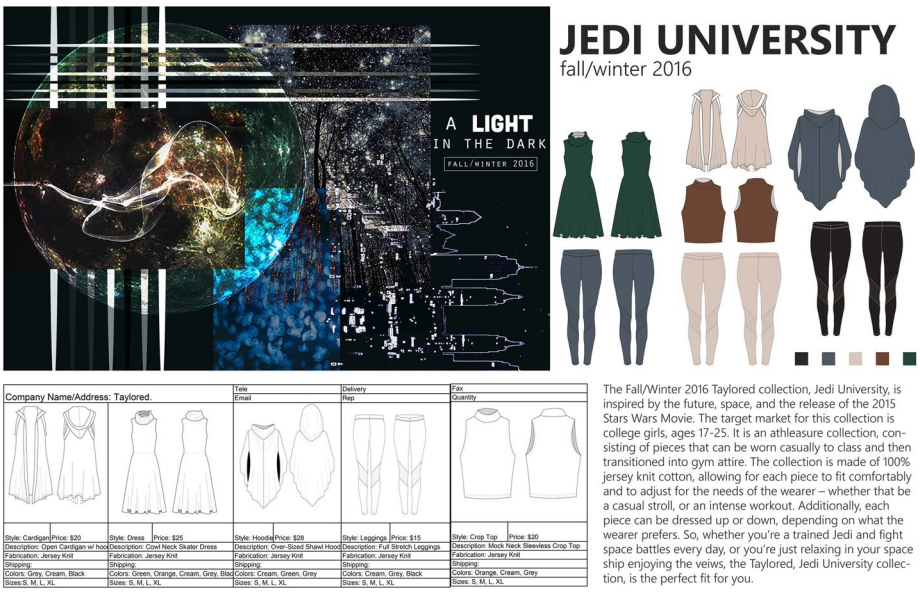

Fig. 3 Example of Category 3 concept board

Fig. 4 Example of Category 4 concept board

and merchandising teams, the presentation or usage boards in this stage connect the product with the overall brand image and identity.

Mood boards encompass various aspects of the fashion industry and contribute significantly at each stage of the product development life cycle. Additionally, mood boards enhance creative thinking and fluency of ideas as well as exploration about the mode and content of the products' visual message. Designers' creative expression and experimentation during the mood board creation allows them to dive deeper into themes or concepts, providing a wealth of information for designers to pull inspiration from. Lastly, through the use of mood boards, designers may be able to communicate visually and effectively, what is a web of seemingly unconnected ideas, difficult to express verbally with similar impacts. In essence, the mood board serves to engage designers in deeper creative thinking, yet evidence is lacking regarding a formalized and reliable assessment instrument to provide critical feedback for mood board creation skill development in academia.

Consensual assessment technique

According to creativity researchers based in psychology and cognitive sciences, there are multiple levels of creativity, with a few recurring thematic ideas (Amabile 1996; Gardner 1993; Runco 2007; Simonton 2009). Generally, a person or product is considered to be creative based on novelty and usefulness to the domain or field of work. In addition to novelty, the creative process and output evaluation and acceptance are commonplace in creativity definitions (Amabile 1983a, 1996; Gardner 1983, 1993; Guilford 1950; Sawyer 2006; Torrance 1962). Early work conducted by Amabile (1983a, b, 1996) developing the Consensual Assessment Technique (CAT) is the foundation for numerous research studies and creativity theory development across domains. The original CAT contained sixteen items evaluating creativity and technical quality. The CAT is well validated in research studies and provides a reliable creativity assessment instrument using a range (expert to novice) of raters (Amabile 1983a, b, 1996; Baer 1997, 1998; Baer et al. 2004; Kaufman et al. 2005, 2007, 2008; Runco 1999). Studies conducted across domains using both adults and children as subjects, have yielded reliability results often exceeding .70 (see Amabile 1996). With early editions, a panel of expert raters independently evaluated parallel products on numerous measures of creativity and technical quality.

Condensed versions (single-item measures) of the CAT have offered similar reliability of creativity assessment, while facilitating the collection of assessment data. For instance, Kaufman et al. (2008) evaluated SciFaiku (Japanese variation on the haiku) poetry written by volunteers. In this study, both novice and expert raters conducted an assessment using a single-item measure asking them to rate the creativity of the poem on a scale of 1–6. Overall, reliability analyses for the separate groups were reported as good (>.80) and excellent (>.90). In a related and similar experiment, Kaufman et al. (2009) used a single question evaluation for creativity to assess short stories, with similar results. Expert and novice film ratings were evaluated using a single-item measure (rating scale 1–10) with success using students, online ratings, and expert film critics (Plucker et al. 2009). Successful results using the CAT are often defined as high levels of interrater reliability (>.80) within and between groups of expert and non-expert raters. This definition is used for the overall assessment of the CAT in the current study. In various domains,

researchers using single-item measures of creativity have reported success using a condensed CAT, while reporting no administration issues, questions of reliability, nor validity concerns.

Expert vs. non-expert raters

In certain domains, would non-experts provide an equally accurate assessment of creativity than experts? While the cost and time of expert panels presents difficulties in creativity assessment, the use of non-experts may not provide a valid assessment of creativity, despite reliability measures (Kaufman and Baer 2012). Early research into the use of experts vs. non-experts indicates the incongruence in responses and creativity assessments; yet with sufficient interrater reliability within groups.

Kaufman et al. (2008) conducted a study involving experts and non-experts in a specified technical domain (poetry) and reported satisfactory to high interrater reliabilities (expert: $\alpha = .83$; non-expert: $\alpha = .94$). However, sample sizes of raters (expert: $n = 10$; non-expert: $n = 106$) effected interrater reliabilities of non-experts; when adjusted for size, non-expert interrater reliabilities dropped significantly to $\alpha = .58$. With little correlation between the groups, conclusions do not support the use or substitution of non-experts in this particular poetry domain. In a less specialized and familiar domain (short-story writing), results using high numbers of non-experts ($n = 100+$) correlated significantly with those of experts ($r = .71$), yet when interrater reliabilities are adjusted for sample size non-expert ratings dropped significantly to $\alpha = .53$ (Kaufman et al. 2009). In repeated studies, using both experts and novices; children and adults; verbal and visual products the CAT provided satisfactorily interrater reliabilities, which usually exceed .70, with some reported as high as the mid .90s. In a majority of these studies, interrater reliabilities typically ranged in the .80s (Amabile 1983a, b, 1996; Baer 1997, 1998; Baer et al. 2004; Kaufman et al. 2005, 2007, 2008; Runco 1999).

In early studies using the CAT to assess creativity levels in collages and drawings of children and adults, Amabile (1983a, 1996) reported inter-rater reliabilities from .70–.90. In subsequent studies focused on image and story creation in children, similar inter-rater reliabilities were reported (Hennessey and Amabile 1999; Runco 1999). In subsequent studies, Kaufman et al. (2005), found gifted novices provided reliable ratings, which were highly correlated with those of expert judges. Plucker et al. (2008) investigated the use of expert, quasi-expert (via professional and user-driven websites), and novice ratings of more than 600 movies. While novice and expert correlations were similar to earlier investigations cited on novice vs. expert ratings, results of the quasi-expert group bridged the gap with significant correlations with both novice ($r = .65$) and experts ($r = .72$). Results indicate a dichotomous classification of raters does not support the full range of options available to researchers for creativity evaluation (Plucker et al. 2008, 2009). Therefore the use of gifted novices or quasi-experts is supported in conjunction with, or as a suitable substitute for, expert ratings of creative products in technical and non-technical domains (Baer et al. 2004, 2009; Kaufman et al. 2005).

Domain specificity of evaluation

Creativity varies across domains, as does the expertise required for valid assessment using the CAT (Baer 1993, 1996). In previously cited studies, evaluation comparisons

between experts, quasi-experts and novices indicated varied results within the domains of poetry, film, writing, etc. While examining the creativity of captions, Kaufman et al. (2007) indicated psychology students maintained adequate interrater reliability and sufficient ratings across writers. Therefore, the use of students evaluating written captions was acceptable.

In an area with no discernable training or apprenticeship, such as mood boards or image captions, can an amateur or a quasi-expert from a related field provide a reliable and valid assessment of creativity? Kaufman and Baer (2012) suggest the use of novice raters for expert-less domains may provide reliable creativity assessment, yet questions of validity would remain of concern. While results from creativity assessment in these domains will produce some validity, further support may be given by comparison between novice and quasi-experts from a related field. Additionally, the use of those involved in the creation or the projects as self and peer assessors may provide background knowledge of the project and increase validity of the creativity assessment. Based on an inversion of the hierarchy of domains (Simonton 2009), the potential or assessment of personal and/or everyday creativity is supported by the through the originality of the idea, even if only for the individual. Based on this premise, in assessing creativity in an expert domain, the use of participant artists/raters seems appropriate with sufficient validity.

Questions linger about the domain specificity of creativity assessment and methods of identifying and classifying experts. Plucker and Runco (1998) stated the uselessness of single predictive creativity measures (e.g. creativity tests), and agreed with Hennessey and Amabile (1988) of using an overall assessment of product output, measured using consensual assessment. In prior research on expert and non-expert creativity assessment, the domains were limited to those ranking higher on Simonton's (2009) hierarchy of domains, such as poetry, and creative writing. Fashion, by its very purpose and use of materials, has hinged between utilitarian artifact with functional appeal to creative applied arts, with exhibitions in the finest museums around the world. With its multiple functions, there exists a deficit in understanding and evaluating creative outputs, starting with the concept mood board stage. In addition, much of the prior research using verbal stimuli resulted in effective creativity evaluation, however assessment of figural stimuli is limited when comparing expert and non-expert assessments of creativity.

Methods

Procedure for mood board assignment

Based on the research intent to consensually assess a creative output, parameters for the assignment were structured in a heuristic method to provide latitude for interpretation, thereby increasing the likelihood of a creative result (Amabile 1983a, b). Students enrolled in second year fashion industry courses were selected to complete a course assignment related to creativity. In addition, participating students served as the non-expert raters for image, except their own. While not a requirement for the course, participant consented to having their work used as part of a research study and to serve as a reviewer. Participants were provided the following instructions:

> *"If someone asks you to take a picture of a flower or a tree that is a pretty straightforward task. But what if we ask you to take a picture of sad? You would have to first*

personalize the word by giving sad a physical feature, and then find an object that represents that feature. For example, a funeral, or a very grey sky, or a heartbroken lover might be the picture you could take to express the abstract meaning of sad Your task is to take five pictures that characterize ONE of the following words of your choice: think, loud, tired, clever, or religious."

The five concept words were selected by the authors based on the vagueness of the term and openness to interpretation by a diverse group of participants. In reviewing a list of possible terms identified through the creativity literature, the researchers narrowed down to the five terms based on fit of the terms to the consensual assessment methodology (Amabile 1983a; Kaufman et al. 2005, 2007). The heuristic nature of the terminology and approach of the instructions were intentional as evidenced by prior creativity researchers when setting task parameters (Amabile 1983a, b, 1996; Baer 1997, 1998; Baer et al. 2004; Kaufman et al. 2005, 2007, 2008; Runco 1999). Photos, selected by the participants to represent their chosen theme, were uploaded onto a single page PDF and formatted to standardize size consistency, with the selected theme as the title heading. Instructors from two different universities asked students in their courses: creative thinking and design and visual design to complete the assignment, resulting in 50 total mood boards. Categorical themes were not even across groups, with students selecting and producing the following mood board themes: clever ($n = 3$), loud ($n = 14$), think ($n = 5$), religious ($n = 16$), and tired ($n = 12$). Figure 5 provides an example of the category 2 style concept boards evaluated by the raters.

Procedure for expert and non-expert raters

After all 50 students completed their mood boards; each file was formatted to maintain consistency for rating and evaluation. Students participated in the project, following IRB consent, based on enrollment in select courses. One of the courses used was

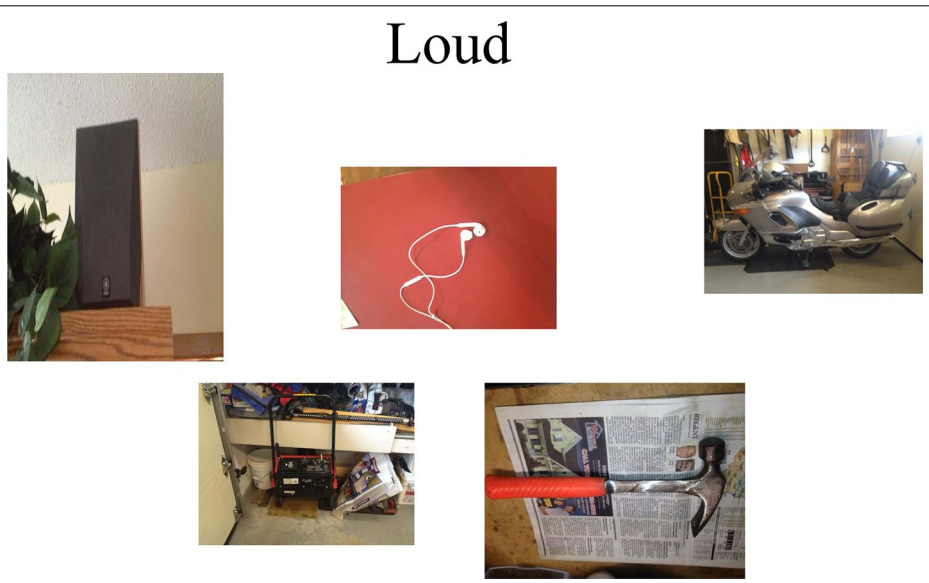

Fig. 5 Example of Category 2 style standardized concept board evaluated by raters

specific to an overview of the fashion industry and is open to all majors. The other course specifically engaged in the concept of creativity as a key overarching theme in the fashion industry, again open to all majors. Both course are open to all classification levels, however most of the students enrolled are a second or third year level. Over 90% of the enrolled students are majoring in fashion design and/or fashion merchandising. The remaining 10% were from various majors in the visual arts. Images were transferred to an electronic survey, where raters were asked to rate the creativity of each collage of images in relation to the theme noted in text. A 5-point Likert scale was used representing the level of creativity for each collage. Ratings ranged form *not creative* (1) to *very creative* (5). Images were randomized for each rater, with raters evaluating all the mood boards. Expert raters ($n = 10$) were selected based on prior academic teaching in the fields of apparel design, visual arts, interior design, and/or architecture. Expert raters ($n = 10$) were purposively sampled and exhibited proficiency in their domain with creative works and scholarship accepted for juried and peer-evaluated national/international competitions. Expert raters were identified as having a knowledge in the area of visual representation of conceptual ideas, specifically fashion design mood boards. 8 of the 10 experts has an educational and/or occupational background in fashion design, while the other two expert raters' educational and/or occupational background is in visual arts. However, both of the non-fashion design experts had worked (occupationally) in textile/fashion design at a stage in their career and exhibited a knowledge of the purpose and intent of conceptual mood boards. Non-expert raters ($n = 88$) included students who completed the projects as well as additional students studying fashion design and/or fashion merchandising across two universities. All 50 students who participated in creating a mood board also rated a mood board, however they did not rate their own mood board. The underlying theory of consensual assessment by non-experts indicates if a significant number of evaluators (>50) are used then individual rater bias is controlled through statistical power (Amabile 1996). Therefore, non-expert rater bias was not a concern for the scope of this project. A minimal amount of extra credit, representing <1% of the total course grade, was provided to students who completed the evaluation and rating.

Data analysis

During initial data analysis and correction, missing ratings were handled using techniques sufficient for data correction. Any rater who did not rate at least 75% of the images was omitted from further data analysis. Of the remaining, none were missing individual ratings for either expert or non-expert raters. Of the 105 non-expert raters who accessed the survey, 101 agreed to continue with the project. A total of 88 of the 101 completed the survey. Of the ten experts to access the survey all ten agreed to participate and fully completed the survey. Consistency among the raters was evaluated using a Cronbach's coefficient alpha. As in many creativity studies coefficient alpha is calculated using the raters as variable items. If there is a difference in the number of raters for experts and non-experts, as is often the case, then an adjusted coefficient alpha should be used for interrater reliability comparison (Kaufman et al. 2008). Due to the size difference between the two groups of raters, $n = 88$ compared to $n = 10$, a Spearman-Brown adjusted coefficient alpha was used to standardize the internal consistency measures

for both groups. In a comparison between coefficient alphas of expert and non-expert raters, confidence intervals were calculated using a standard error. Alpha level of .05 was set for the difference. An independent samples t test was calculated to compare means between the two groups for significant differences. Confident that appropriate assumptions for correlational analysis were met, a Pearson correlation between expert and non-expert ratings was conducted to assess the relationship between two group ratings.

A repeated measure of analysis (ANOVA) was computed to examine the differences among the creativity ratings' means of both non-expert and expert raters and the subject matter of each of the mood boards (clever, loud, think, religious, and tired). Data analysis included examination of significant differences between subject matter categories for each group (non-expert and expert) as well as an overall assessment for all raters.

Results and discussion

A total of 88 non-expert and 10 expert raters evaluated all 50-mood boards. Mean rating score for non-experts was $M = 2.83$, $SD = .32$, while results reported from expert ratings was $M = 3.26$, $SD = .37$. Independent samples t test comparing group means indicate expert raters evaluated the mood boards significantly more creative than the non-experts, $t\ (99) = -6.71$, $p < .001$, (95% CI −.57 to −.29). Pearson correlation results indicated a significant relationship between the two groups of raters, $r\ (50) = .33$, $p < .01$. The effect size was between medium and large (Cohen 1994). Interrater reliability analysis of expert raters indicated an insufficient alpha level, $\alpha = .66$ (95% CI .50–.79). Reliability analysis of expert ratings fell just outside of an acceptable cutoff ($\alpha > .70$) (Gliner et al. 2002). Conversely, non-expert rater reliability results indicated high levels of interrater agreement, typical of large groups using the CAT, $\alpha = .92$ (95% CI .88–.95). When adjusting for sample size, results of the Spearman-Brown adjusted coefficient alpha formula indicated a drop in overall reliability. For the expert raters, coefficient alphas dropped to $\alpha = .49$, well below the threshold for sufficient analysis of reliability. While the non-expert ratings also decreased ($\alpha = .86$), reliability results remained at a sufficient to high level of acceptability. As seen in earlier studies (Kaufman et al. 2008), adjusting the coefficient alphas, using Spearman-Brown formula, provided a stronger basis for comparison, even when results from an expert-less domain differ from those analyzing domains such as poetry.

Categorical themes were not even across groups: clever ($n = 3$), loud ($n = 14$), think ($n = 5$), religious ($n = 16$), and tired ($n = 12$); therefore, an overall mean creativity ratings was used. ANOVA results for all raters indicated a significant difference between the five subject matter categories; $F\ (4, 95) = 4.64$, $p < .005$. Clever received the highest creativity ratings $M = 3.51$, which may be attributed to the limited number of projects under this category. Yet, *think*, with a relatively smaller representation ($n = 5$) reported the lowest creativity ratings of the five categories $M = 2.77$. Post hoc Games–Howell tests indicated significant differences between the categories *think* and *tired* ($p = .03$, $d = 1.24$). Comparison of the remaining categories revealed no significant difference between the subject matter and creativity ratings for all raters.

ANOVA results for the non-expert raters similarly reflected those of the overall group assessment, which was anticipated considering the sample size of the non-experts compared to experts. An overall significant difference was reported between the groups, F

(4, 45) = 6.39, $p < .001$. However, the assumption of equal variances was violated and post hoc Games–Howell results indicated significant differences between *tired* and *loud* ($p = .04$, $d = 1.19$) as well as *tired* and *religious*, ($p = .02$, $d = 1.27$). Therefore, compared with overall ratings from both non-experts and experts, non-experts indicated significant differences between the subject matter of three of the categories with similar sample sizes. Expert raters' overall ANOVA results indicated significant differences between the groups as well, $F(4, 45) = 5.22$, $p = .002$, however assumption of equal variances were not violated. Post hoc analysis included Tukey HSD and indicated significant differences between *think* and *clever* ($p = .009$, $d = 2.13$), *think* and *religious* ($p = .002$, $d = 1.93$), and *think* and *tired* ($p = .036$, $d = 1.45$). Results suggest a significant difference between assessments of all raters' creativity evaluations, namely between the *think* and *tired* categories. When broken down to non-experts and experts, categorical differences varied between the two groups. Non-experts reported significant results namely with categories compared with *tired*, while experts reported significant differences with categories reported against *think* (Table 1). Complete post hoc analysis of the combined raters two category comparisons are included in Table 2.

The analysis of the themed mood boards focused on the creativity of content selected. As previously mentioned, the higher the incongruity and un-relatedness of imagery, the more depth a designer has to pull from for inspiration. Based on the content provided and assessed, religious themes were extremely prevalent explained perhaps by the ease and familiarity of the concept with most people. Using a single-item measure for creativity assessment and evaluation facilitated large amounts of feedback within a short amount of time. For this investigation, experts assembled were faculty with experience evaluating and creating photo mood boards, namely those which fit into a thematic response, similar to trend boards and/or artistic mood boards. Additionally, when creative products are displayed digitally and formatted similarly, this study supported the use of digitization and standardization of image size, which may have affected the interrater reliability of experts [$\alpha = .66$ (95% CI .50–.79)], when compared to previous research documenting the use of the CAT (Baer et al. 2004, 2009; Kaufman et al. 2005). Conversely, non-expert evaluators reported consistently high levels of reliability, even when adjusted for group size, yet the validity of creativity assessment is compromised by the lack of individual training, exposure, and experience within the domain. Therefore, expert raters using the single-item digital measure, while providing an assumption of validity, lacked overall reliability and agreement between raters. These results are not

Table 1 One-way analysis of variance (ANOVA) and means comparison for consensual assessment technique (combined, non-expert, and expert raters)

Raters	Clever (n = 6), *M*	Loud (n = 28), *M*	Religious (n = 32), *M*	Think (n = 10), *M*	Tired (n = 24), *M*	*F*	*P*	(*df*)[a]
Combined	3.51	2.93	3.07	2.77	3.14	4.64	.005	(4, 95)*
Non-expert	3.42	2.72	2.72	2.76	3.00	6.39	.001	(4, 45)***
Expert	3.60	3.14	3.43	2.78	3.29	5.22	.002	(4, 45)**

Each image was rated a five-point scale, ranging from "Not very creative" (1) to "Very creative" (5)

* $p < .05$, ** $p < .01$, *** $p < .001$

[a] Scattered missing values account for variability in degrees of freedom for each item

Table 2 One-way analysis of variance (ANOVA) Tukey HSD post hoc (combined raters)

Categories	Mean difference	Significance
Clever		
Loud	.57*	.009
Religious	.43	.083
Think	.73*	.003
Tired	.36	.235
Loud		
Clever	−.57*	.009
Religious	−.14	.597
Think	.15	.790
Tired	−.21	.251
Religious		
Clever	−.43	.083
Loud	.14	.597
Think	.30	.194
Tired	−.07	.951
Think		
Clever	−.73*	.003
Loud	−.15	.790
Religious	−.30	.194
Tired	−.37	.074
Tired		
Clever	−.36	.235
Loud	.21	.251
Religious	.07	.951
Think	.37	.074

Note * $p < .01$

supported by the literature, which in prior studies indicated expert raters to exhibit the highest levels of interrater reliability (Plucker et al. 2008, 2009). However, non-expert evaluators tended to highly agree between raters, but it remains unclear as to what exactly they are evaluating and why. Non-expert results support prior research both in interrater agreement and comparison to expert raters for consensual assessment (Baer et al. 2004, 2009; Kaufman et al. 2005). Based on these results, it is recommended to use both expert and non-expert raters to provide a comprehensive assessment which is both reliable and valid.

Implications

As creativity assessment moves across domains, this question regarding qualifications of judges will provide additional questions about the reliability and validity of consensual assessment. Prior research using consensual assessment, indicated substituting non-experts for experts was not recommended, and this investigation continues to support that position. Although, ideally combining the use of expert and non-expert raters, provides reliable assessments and validity support. Maintaining equal group sizes is recommended. Not only will students be provided with reliable feedback on creativity of content using this measure, but they will receive assessment from a variety of sources

and backgrounds. By seeking more substantive feedback, students can focus mood board creation skill development on specific deficiencies. However, the first step in development and progress is a true and reliable assessment of current conditions. The ease of use of the single digital measure to evaluate creativity of visual stimuli provides academics and students an initial assessment of the quality of content selected prior to subjective aesthetic judgment.

Objectivity is becoming the norm in the current education grading systems. However, in disciplines such as apparel design or the applied arts, objectivity is often difficult to achieve and may lead the student to discredit creativity evaluation as one person's opinion. Therefore, through the use of digital consensual assessment, the subjective expert opinion of a faculty expert is supported by the assessment of non-experts with an interest in the field (peers). This procedure will enable effective multi-channel feedback for the designer/artist and provide unique learning opportunities in receiving and reacting to critical assessments. While this tool will not fully replace the keen expert eye of a trained educator, use of this instrument will go a long way in helping to support the subjective assessment of creativity in academia.

Limitations of this study include unequal distribution of mood boards under each of the five categories. However, the effectiveness of using a single item digital measure for creativity evaluation is not diminished when compared with previous studies using consensual assessment. Additionally, the difference in group size between expert and non-expert raters is a consistent issue from the literature. While it would be ideal from a statistical analysis perspective to maintain equal groups, the difficulty in amassing a panel of experts is often a hindrance. Further examination to measure the reliability and validity of the instrument with equal sizes is needed and recommended for future studies. From an education and research perspective, the use of this instrument can provide an efficient and simple creativity evaluation measure with adequate reliability, continuing to promote the research and teaching of creativity with quantitative evaluation. Further studies examining variation in mood board styles and features as well as depth of assessment will provide additional scholarship contributing to the field. Additionally, identifying mood board creation skills will enable assessment research to target and evaluate specific skills for further enhancement. The use of mood boards to express the creativity of inspiration digitally will continue to grow in the fashion industry. Therefore, continued research evaluating, critiquing, and enhancing digital mood boards and creative expression will be necessary.

Authors' contributions
CF conducted the consensual assessment data collection and drafted the initial manuscript. SM drafted the assignment and collected data from participants, as well as served as a reviewer of consensual assessment. EK drafted the assignment and conducted results analysis and interpretation as well as served as a reviewer of consensual assessment. All authors read and approved the final manuscript.

Author details
[1] Mississippi State University, Mississippi State, MS, USA. [2] Iowa State University, Ames, IA, USA.

Competing interests
The authors declare that they have no competing interests.

Received: 21 December 2016 Accepted: 1 May 2017
Published online: 28 August 2017

References

Amabile, T. (1983a). *The social psychology of creativity*. New York: Springer.
Amabile, T. (1983b). The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology, 45,* 357–376.
Amabile, T. (1996). *Creativity in context*. Boulder: Westview Press.
Baer, J. (1993). *Creativity and divergent thinking: A task-specific approach*. Presented in the Berlyne Prize address at the annual meeting of the American Psychological Association, Washington, DC, August 1992.
Baer, J. (1996). The effects of task-specific divergent-thinking training. *The Journal of Creative Behavior, 30*(3), 183–187.
Baer, J. (1997). Gender differences in the effects of anticipated evaluation on creativity. *Creativity Research Journal, 10,* 25–31.
Baer, J. (1998). The case for domain specificity of creativity. *Creativity Research Journal, 11*, 173–177.
Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*(1), 113–117.
Baer, J., Kaufman, J. C., & Riggs, M. (2009). Brief report: Rater-domain interactions in the consensual assessment technique. *The International Journal of Creativity and Problem Solving, 19*(2), 87–92.
Boyes, J. (1998). Essential fashion design: illustration, theme boards, body coverings, projects, portfolios. London: Batsford.
Cassidy, T. D. (2008). Mood boards: Current practice in learning and teaching strategies and students' understanding of the process. *International Journal of Fashion Design, 1*(1), 43–54.
Cassidy, T. (2011). The mood board process modeled and understood as a qualitative design research tool. *Fashion Practice, 3*(2), 225–251.
Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.
Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences* (10th Anniv ed.). New York: Basic Books.
Gardner, H. (1993). *Creating minds: An anatomy of creativity seen through the lives of Freud, Einstein, Picasso, Stravinsky, Eliot, and Gandhi*. New York: Basic Books.
Garner, S., & McDonagh-Philp, D. (2001). Problem interpretation and resolution via visual stimuli: The use of 'mood boards' in design education. *Journal of Art and Design Education, 20*(1), 57–64.
Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): what do the textbooks say? *The Journal of Experimental Education*, *71*(1), 83–92.
Guilford, J. P. (1950). Creativity. *American Psychologist, 5,* 444–454.
Hennessey, B. A., & Amabile, T. M. (1988). Storytelling: A method for assessing children's creativity. *Journal of Creative Behavior, 22,* 235–246.
Hennessey, B. A., & Amabile, T. M. (1999). Consensual assessment. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (pp. 347–359). San Diego: Academic Press.
Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal, 24*(1), 83–91.
Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior, 43*(4), 223–233.
Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*(2), 171–178.
Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way? *Gifted Child Quarterly, 49*(3), 260–265.
Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity, 2*(2), 96–106.
Lucero, A. (2012). Framing, aligning, paradoxing, abstracting, and directing: How design mood boards work. In *Proceedings of the designing interactive systems conference,* pp. 438–447.
McDonagh, D., & Storer, I. (2004). Mood boards as a design catalyst and resource: Research an under-researched area. *The Design Journal, 7*(3), 16–31.
Plucker, J. A., Holden, J., & Neustadter, D. (2008). The criterion problem and creativity in film: Psychometric characteristics of various measures. *Psychology of Aesthetics, Creativity, and the Arts, 2*(4), 190.
Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way? *Psychology and Marketing, 26*(5), 470–478.
Plucker, J., & Runco, M. (1998). The death of creativity measurement has been greatly exaggerated: Current issues, recent advances, and future directions in creativity assessment. *Roeper Review, 21,* 36–39.
Runco, M. A. (1999). A longitudinal study of exceptional giftedness and creativity. *Creativity Research Journal, 12,* 161–164.
Runco, M. A. (2007). *Creativity theories and themes: Research, development and practice*. Burlington: Elsevier Academic Press.
Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation*. New York: Oxford University Press.
Simonton, D. K. (2009). Varieties of perspectives on creativity reply to commentators. *Perspectives on Psychological Science, 4*(5), 466–467.
Torrance, E. P. (1962). *Guiding creative talent*. Englewood Cliffs: Prentice-Hall.